

Extending Weakly-Sticky Datalog[±]: Query-Answering Tractability and Optimizations

Mostafa Milani and Leopoldo Bertossi

Carleton University, School of Computer Science
Ottawa, Canada

Abstract. *Weakly-sticky* (*WS*) Datalog[±] is an expressive member of the family of Datalog[±] programs that is based on the syntactic notions of *stickiness* and *weak-acyclicity*. Query answering over the *WS* programs has been investigated, but there is still much work to do on the design and implementation of practical query answering (QA) algorithms and their optimizations. Here, we study sticky and *WS* programs from the point of view of the behavior of the chase procedure, extending the stickiness property of the chase to that of *generalized stickiness of the chase* (*gscH-property*). With this property we specify the semantic class of *GSCH* programs, which includes sticky and *WS* programs, and other syntactic subclasses that we identify. In particular, we introduce *joint-weakly-sticky* (*JWS*) programs, that include *WS* programs. We also propose a bottom-up QA algorithm for a range of subclasses of *GSCH*. The algorithm runs in polynomial time (in data) for *JWS* programs. Unlike the *WS* class, *JWS* is closed under a general magic-sets rewriting procedure for the optimization of programs with existential rules. We apply the magic-sets rewriting in combination with the proposed QA algorithm for the optimization of QA over *JWS* programs.

1 Introduction

Ontology-based data access (OBDA) [24] allows to access, through a conceptual layer that takes the form of an ontology, underlying data that is usually stored in a relational database. Queries can be expressed in terms of the ontology language, but are answered by eventually appealing to the extensional data underneath. Common languages of choice for representing ontologies are certain classes (or fragments) of *description logic* (DL) [3] and, more recently, of *Datalog[±]* [8, 10]. Those classes are expected to be computationally well-behaved in relation to query answering (QA). Several approaches for QA, and a number of techniques have been proposed for DL-based [3, 24] and Datalog[±]-based OBDA [8]. In this work we concentrate on the conjunctive QA problem from relational data through Datalog[±] ontologies.

Datalog[±], as an extension of the Datalog query language [11], allows in rule heads (i.e. consequents): existentially quantified variables (\exists -variables), equality atoms, and a false propositional atom, say **false**, to represent “negative program constraints” [8–10]. Hence the “+” in Datalog[±], while the “−” reflects syntactic restrictions on programs for better computational properties.

Datalog[±] is expressive enough to represent in logical and declarative terms useful ontologies, in particular those that capture and extend the common conceptual data models [9] and Semantic Web data [2]. The rules of a Datalog[±] program can be seen as forming an ontology on top of an extensional database, D , which may be *incomplete*. In particular, the ontology: (a) provides a “query layer” for D , enabling OBDA, and (b) specifies a completion of D .

In the rest of this work we will assume that programs contain only existential rules (plus extensional data). When programs are subject to syntactic restrictions, we talk about Datalog[±] programs, whereas when no conditions are assumed or applied, we talk about Datalog⁺ programs, also called Datalog[∃] programs [4, 8, 15, 16].

From the semantic and computational point of view, the completion of the underlying extensional instance D appeals to so-called *chase* procedure that, starting from D , iteratively enforces the rules in the ontology. That is, when a rule body (the antecedent) becomes true in the instance so far, but not the head (the consequent), a new tuple is generated. This process may create new values (nulls) or propagate values to the same or other *positions*. The latter correspond to the arguments in the schema predicates.

Example 1. Consider a Datalog[±] program \mathcal{P} with extensional database $D = \{r(a, b)\}$ and set of rules \mathcal{P}^r :

$$r(X, Y) \rightarrow \exists Z \, r(Y, Z). \quad (1) \quad r(X, Y), r(Y, Z) \rightarrow s(X, Y, Z). \quad (2)$$

The positions for this schema are: $r[1], r[2], s[1], s[2], s[3]$. The extension of D generated by the chase includes the following tuples (among infinitely many others): $r(b, \zeta_1), s(a, b, \zeta_1), r(\zeta_1, \zeta_2), s(b, \zeta_2, \zeta_1)$. Notice that $s(a, b, \zeta_1)$ and $s(b, \zeta_1, \zeta_2)$ are obtained by replacing the *join variable* Y (i.e. repeated) in the body of (2) by b and ζ_1 , resp. ■

The result of the chase, seen as an instance for the combined ontological and relational schema, is also called “the chase”. The chase (instance) extends D , but may be infinite; and gives the semantics to the Datalog[±] ontology, by providing an intended model, and can be used for QA. At least conceptually, the query can be posed directly to the materialized chase instance. However, this may not be the best way to go about QA, and computationally better alternatives have to be explored.

Actually, when the chase is infinite, (conjunctive) QA may be undecidable [14]. However, in some cases, even with an infinite chase, QA is still computable (decidable), and even tractable in the size of D . In fact, syntactically restricted subclasses of Datalog⁺ programs have been identified and characterized for which QA is decidable, among them: *linear*, *guarded* and *weakly-guarded*, *sticky* and *weakly-sticky* (*WS*) [8, 10] Datalog[±].

Sticky Datalog[±] is a syntactic class of programs characterized by syntactic restrictions on join variables. *WS* Datalog[±] extends sticky Datalog[±] by also capturing the well-known class of *weakly-acyclic programs* [13], which is defined in terms of the syntactic notions of *finite-* and *infinite-rank* positions. Accordingly, *WS* Datalog[±] is characterized by restrictions on join variables occurring

in infinite-rank positions. A non-deterministic QA algorithm for WS Datalog $^\pm$ is presented in [10], to establish the theoretical result that QA can be done in polynomial-time in data.

In this work, we concentrate on sticky and WS Datalog $^\pm$, because they have found natural applications in our previous work on extraction of quality data from possible dirty databases [20]. The latter task is accomplished through QA, so that the need for efficient QA algorithms becomes crucial. Accordingly, the main motivations, goals, and results (among others) for/in this work are:

- (A) Providing a practical, bottom-up QA algorithm for WS Datalog $^\pm$. Being bottom-up, it is expected to be based on (a variant of) the chase. Since the latter can be infinite, the query at hand guarantees that the need to generate only an initial, finite portion of the chase.
- (B) Optimizing the QA algorithm through a *magic-sets* rewriting technique, to make it more query sensitive.

For (B), we apply the magic-sets technique for Datalog $^+$ first introduced in [1], which we denote with \mathbf{MagicD}^+ . Extending classical magic-sets for Datalog [11], \mathbf{MagicD}^+ prevents existential variables from getting bounded, a reasonable adjustment that essentially preserves the semantics of existential rules during the rewriting. Unfortunately, the class of WS Datalog $^\pm$ programs is provably not closed under \mathbf{MagicD}^+ , meaning that the result of applying \mathbf{MagicD}^+ to a WS program may not be WS anymore. This led us to search for a more general class of programs that is: (i) closed under \mathbf{MagicD}^+ , (ii) extends WS Datalog $^\pm$, and (iii) has an efficient QA algorithm. Notice that at this point both syntactic and semantic classes may be investigated, and we do so. The latter classes refer to the properties of the chase as an instance.

Sticky programs enjoy the *stickiness property of the chase*, which -in informal terms- means the following: If, due to the application of a rule during the chase, a value replaces a join variable in the rule body, then that value is propagated through all the possible subsequent steps, i.e. the value “sticks”. The “stickiness property of the chase” defines a “semantic class”, SCh , in the sense that it is characterized in terms of the chase for programs that include an extensional database. This class properly extends sticky Datalog $^\pm$ [10].

We can relax the condition in the *sch-property*, and define the *generalized-stickiness property of the chase*. It is as for the *sch-property*, but with the propagation condition only on join variables that *do not* appear in the *finite positions*; the latter being those where finitely many different values may appear during the chase. With this property we define the new semantic class of $GSCh$ programs. However, we make notice that, given a program \mathcal{P} consisting of a set of rules \mathcal{P}^r and an extensional instance D , computing (deciding) $FinPoss(\mathcal{P})$, the set of finite positions of \mathcal{P} , is unsolvable (undecidable) [12]. Accordingly, it is also undecidable if a Datalog $^+$ program belongs to the $GSCh$ class.

Starting from the definition of the $GSCh$ class, we can define, backwardly, a whole range of different semantic classes between *Sticky* and $GSCh$, by replacing in the definition of the latter the condition on the set of non-finite positions by a stronger one that appeals to a superset of them. Each of these supersets is

represented through its complement, which is determined by an abstract *selection function* \mathcal{S} that identifies a set of finite positions. Such a function, given a program \mathcal{P} , returns a subset $\mathcal{S}(\mathcal{P})$ of $FinPoss(\mathcal{P})$ (making \mathcal{S} sound, but possibly incomplete w.r.t $FinPoss(\mathcal{P})$). \mathcal{S} may be computable or not, and may depend on \mathcal{P}^r alone or on the combination of \mathcal{P}^r and D . Hence we split \mathcal{P} into \mathcal{P}^r and D . The corresponding semantic class of programs, those enjoying the \mathcal{S} -stickiness property of the chase, is denoted with $SCh(\mathcal{S})$.

In particular, if \mathcal{S}^\top is the non-computable function that selects all finite positions, $GSCh = SCh(\mathcal{S}^\top)$. If \mathcal{S}^{rank} selects the finite-rank positions (that happen to be finite positions) [13], then $WSCh = SCh(\mathcal{S}^{rank})$ is a new semantic class programs, those with the *weak-stickiness property of the chase*. And for the class SCh of programs we started from above, it holds $SCh = SCh(\mathcal{S}^\perp)$, with \mathcal{S}^\perp always returning the empty set of positions. Notice that \mathcal{S}^{rank} and \mathcal{S}^\perp are both computable, and they do not use the extensional instance D , but only the program. In this sense, we say that they are *syntactic selection functions*.

We can see that the combination of selection functions with the \mathcal{S} -based notion of stickiness property of the chase (i.e. that only values in join variables in positions outside those selected by \mathcal{S} propagate all the way through), defines a range of semantic classes of programs starting with SCh , ending with $GSCh$, and with $SCh(\mathcal{S}^{rank})$ in between. They are shown in ascending order of inclusion, from left to right, in the middle layer of Figure 1. There, the upper layer shows the corresponding selection functions ordered by inclusion (of their images).

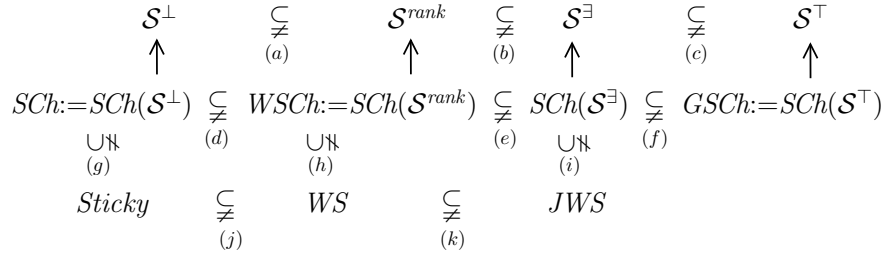


Fig. 1: Semantic and syntactic program classes, and selection functions

A parallel and corresponding range of syntactic classes, also ordered by set inclusion, is shown in the lower layer. It includes the sticky and WS classes (cf. Figure 1, bottom). Each syntactic class only partially represents its semantic counterpart, in the sense that the former: does not consider extensional instances, appeals to the same selection function, but also imposes additional syntactic conditions on the set of rules. All the inclusions in Figure 1 are proper, as examples we provide in this work will show (but (g) and (j) are known [10]).

In this work, our main goal is to introduce and investigate the semantic class $SCh(\mathcal{S}^\exists)$, determined by the selection function \mathcal{S}^\exists that is defined in terms of the *existential dependency graph* of a program [15] (a syntactic, computable construction). We also introduce and investigate its corresponding syntactic class of *joint-weakly-sticky* (JWS) programs. The latter happens to satisfy desiderata (A) and (B) above. Actually, about (A), we provide for the class $SCh(\mathcal{S}^\exists)$ a polynomial-time, chase-based, bottom-up QA algorithm, which can be applied

to *JWS* (and all its semantic and syntactic subclasses) in particular. This is a general situation: The polynomial-time QA algorithms for the classes *Sticky* [10], *WS* [10,22], and *JWS* (this work) rely basically on the properties of the semantic class rather than on the specific syntactic restrictions. Hence our interest is in investigating the particular semantics classes, and semantic classes in general, as defined by selection functions. About (B), notice that if we start with a *WS* program, we can apply **MagicD**⁺ to it, obtaining a *JWS* program, for which QA can be done in polynomial time.

The paper is structured as follows: Section 2 is a review of some basics of the database theory, the chase procedure, and Datalog[±]. Section 3 contains the definition of the stickiness and general-stickiness properties of the chase and the *SCh* and *GSCh* semantic classes. Section 4 is about the ranges of syntactic and semantic program subclasses of *GSCh*. The *JWS* class of programs is introduced in Section 5. Section 6 and Section 7 contain the QA algorithm and **MagicD**⁺. In this paper we use mainly intuitive and informal introductions of concepts and techniques, illustrated by examples. The precise technical developments can be found in the Appendices of [23].

2 Preliminaries

We start with a relational schema \mathcal{R} containing two disjoint “data” sets: \mathcal{C} , a possibly infinite domain of *constants*, and \mathcal{N} , of infinitely many *labeled nulls*. It also contains predicates of fixed and finite arities. If p is an n -ary predicate (i.e. with n arguments) and $1 \leq i \leq n$, $p[i]$ denotes its i -th position. With \mathcal{R} , \mathcal{C} , \mathcal{N} we can build a language \mathcal{L} of first-order (FO) predicate logic, that has \mathcal{V} as its infinite set of *variables*. We denote with \bar{X} , etc., finite sequences of variables. A *term* of the language is a constant, a null, or a variable. An *atom* is of the form $p(t_1, \dots, t_n)$, with $p \in \mathcal{R}$, n -ary predicate, and t_1, \dots, t_n terms. An atom is *ground*, if it contains no variables. An *instance* I for schema \mathcal{R} is a possibly infinite set of ground atoms. The *active domain* of an instance I , denoted $Adom(I)$, is the set of constants or nulls that appear in I . Instances can be used as interpretation structures for the FO language \mathcal{L} . Accordingly, we can use the notion of formula satisfaction of FO predicate logic.

A conjunctive query (CQ) is a FO formula, $\mathcal{Q}(\bar{X})$, of the form: $\exists \bar{Y} (p_1(\bar{X}_1) \wedge \dots \wedge p_n(\bar{X}_n))$, with $\bar{Y} := (\bigcup \bar{X}_i) \setminus \bar{X}$. For an instance I , $\bar{t} \in (\mathcal{C} \cup \mathcal{N})^n$ is an *answer* to \mathcal{Q} if $I \models \mathcal{Q}[\bar{t}]$, with \bar{t} replacing the variables in \bar{X} . $\mathcal{Q}(I)$ denotes the set of answers to \mathcal{Q} in I . \mathcal{Q} is Boolean (a BCQ) when \bar{X} is empty, and when true in I , $\mathcal{Q}(I) := \{yes\}$. Otherwise, $\mathcal{Q}(I) = \emptyset$. Notice that a CQ can be expressed as a rule of the form $p_1(\bar{X}_1), \dots, p_n(\bar{X}_n) \rightarrow ans_{\mathcal{Q}}(\bar{X})$, where $ans_{\mathcal{Q}}(\cdot) \notin \mathcal{R}$ is an auxiliary predicate. The query answers form the extension of the answer-collecting predicate $ans_{\mathcal{Q}}(\cdot)$.¹

A *tuple-generating dependency* (TGD), also called *existential rule* or simply a *rule* is a sentence, σ , of \mathcal{L} of the form: $p_1(\bar{X}_1), \dots, p_n(\bar{X}_n) \rightarrow \exists \bar{Y} p(\bar{X}, \bar{Y})$, with \bar{X}_i indicating the variables appearing in p_i (among possibly elements from

¹ When \mathcal{Q} is Boolean, $ans_{\mathcal{Q}}$ is a propositional atom; and if \mathcal{Q} is true in I , then $ans_{\mathcal{Q}}$ can be reinterpreted as the query answer.

\mathcal{C}), and an implicit universal quantification over all variables in $\bar{X}_1, \dots, \bar{X}_n, \bar{X}$, and $\bar{X} \subseteq \bigcup_i \bar{X}_i$, and the dots in the antecedent standing for conjunctions.² The variables in \bar{Y} , that could be empty, are *existential variables*. With $head(\sigma)$ and $body(\sigma)$ we denote the sets of atoms in the consequent and the antecedent of σ , respectively. The notions of satisfaction by an instance I of a TGD σ (denoted $I \models \sigma$), and of a set of TGDs, are defined as in FO logic.

A Datalog⁺ program \mathcal{P} consists of a set of rules \mathcal{P}^r and an extensional database instance D , i.e. a finite instance whose atoms contain only elements from \mathcal{C} . The set of models of \mathcal{P} , denoted by $Mod(\mathcal{P})$, contains all instances I , such that $I \supseteq D$ and $I \models \mathcal{P}^r$. Given a CQ \mathcal{Q} , the set of answers to \mathcal{Q} from \mathcal{P} is defined by $ans(\mathcal{Q}, \mathcal{P}) := \bigcap_{I \in Mod(\mathcal{P})} \mathcal{Q}(I)$.

The *chase* procedure is a fundamental algorithm in different database problems, including implication of database dependencies, query containment, and CQ answering under dependencies [6,10,13,14,17]. For the latter problem [10,13], the idea is that, given a set of dependencies over a database schema and an instance as input, the chase enforces the dependencies by adding new tuples into the instance, so that the result satisfies the constraints (cf. Appendix B in [23] for more details).

Example 2. (example 1 cont.) With the given instance D and the assignment $\theta: X \mapsto a, Y \mapsto b$, rule (1) is not satisfied: $D \models r(X, Y)[\theta]$, but $D \not\models \exists Z r(Y, Z)[\theta]$. Then, the chase inserts a new tuple $r(b, \zeta_1)$ into D (ζ_1 is a fresh null), resulting in instance D_1 . D_1 does not satisfy (2), so the chase inserts $s(a, b, \zeta_1)$, resulting in instance D_2 . The chase continues, without stopping, creating an infinite instance: $chase(\mathcal{P}) = \{r(a, b), r(b, \zeta_1), s(a, b, \zeta_1), r(b, \zeta_1), r(\zeta_1, \zeta_2), s(b, \zeta_1, \zeta_2), \dots\}$. ■

The instance resulting from the chase procedure is also called “the chase”. As such, it is a so-called *universal model* [13], i.e. a representative of all models in $Mod(\mathcal{P})$. In particular, the answers to a CQ \mathcal{Q} under \mathcal{P} , i.e. those in $ans(\mathcal{Q}, \mathcal{P})$, can be computed by evaluating \mathcal{Q} over the chase (and discarding the answers containing nulls). The chase procedure may not terminate, and it is in general undecidable if it terminates, even for a fixed instance [12].

Several sufficient conditions, syntactic [12,13,18] and data-dependent [19], that guarantee chase termination have been identified. *Weak-acyclicity* [13] is one of the former, and is defined using the dependency graph.

Example 3. (example 2 cont.) The *dependency graph* (DG) of \mathcal{P}^r (cf. Figure 2) is a directed graph whose vertices are the positions of \mathcal{R} .

The edges are defined as follows: for every $\sigma \in \mathcal{P}^r$, \forall -variable X in $head(\sigma)$, and position π in $body(\sigma)$: 1. for each occurrence of X in position π' in $head(\sigma)$, create an edge from π to π' . 2. for each \exists -variable Z in position π'' in $head(\sigma)$, create a *special edge* (dashed) from π to π'' .

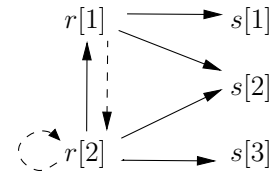


Fig. 2: Dependency graph

² A query of this form can be seen and treated as a new TGD containing a fresh head predicate.

The *rank of a position* is the maximum number of special edges over all (finite or infinite) paths ending at that position. $\Pi_F(\mathcal{P}^r)$ is the set of finite-rank positions in \mathcal{P}^r . A program is *weakly-acyclic* (WA) if all of the positions have finite-rank. Here, $r[1], r[2] \notin \Pi_F(\mathcal{P}^r)$, so the program is not WA. ■

In a program with finite- and infinite-rank positions, every finite-rank position is finite: For any extensional instance D , during the chase only polynomially many different values appear in them (in data) [10]. However, in infinite-rank positions, there may be infinitely many values (and the chase does not terminate). In particular, for every WA program and instance D the chase terminates in polynomially many steps with respect to the size of D [13].

The notions of finite and infinite positions mentioned above rely on the chase instance and hence a program's data: Given a program \mathcal{P} with schema \mathcal{R} , the set of finite positions of \mathcal{P} , that we refer to as $\text{FinPoss}(\mathcal{P})$, is the set of positions where finitely many values appear in $\text{chase}(\mathcal{P})$. Every position that is not finite is infinite.

Conjunctive query answering w.r.t an arbitrary set of TGDs is in general undecidable [5]. The Datalog[±] family is formed by syntactic subclasses of Datalog⁺ programs that are defined by imposing restrictions on the sets of TGDs rules in a program, to guarantee decidability, and in several cases, tractability of QA. In this work we concentrate on the sticky and WS classes of programs.

3 Stickiness of the Chase and its Generalization

The “*stickiness property of the chase*” (*sch-property*) [10] is a “semantic” property of Datalog⁺ programs in relation to the way the chase behaves with the extensional data. We informally introduce it here. A program has this property if, due to the application of a rule σ , when a value replaces a repeated variable in a rule-body, then that value also appears in all the head atoms obtained through the iterative enforcement of applicable rules that starts with σ 's application. In short, the value is propagated through all possible subsequent chase steps.

Example 4. Consider \mathcal{P}_1 with $D_1 = \{r(a, b), r(b, c)\}$, and \mathcal{P}_1^r containing:

$$r(X, Y), r(Y, Z) \rightarrow p(Y, Z). \quad p(X, Y) \rightarrow \exists Z s(X, Y, Z). \quad s(X, Y, Z) \rightarrow u(Y).$$

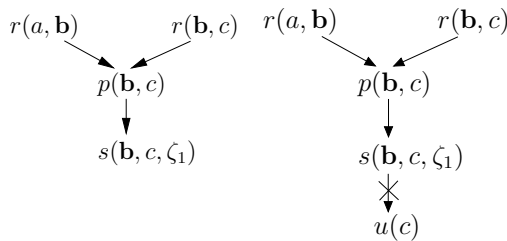


Fig. 3: The *sch-property*.

\mathcal{P}_1 does not have the *sch-property*, as the chase in Figure 3 (right-hand side) shows: value b is not propagated all the way down to $u(c)$. However, a program \mathcal{P}_2 with the same database $D_2 = D_1$ but a set \mathcal{P}_2^r of rules which is \mathcal{P}_1^r without its third rule, has the *sch-property*, as shown in Figure 3 (left-hand side). ■

SCh is the semantic class of programs with the *sch-property*. Next, we briefly recall the classes of programs whose definitions are related to the *sch-property* and the *SCh* programs.

Sticky Programs. Sticky Datalog[±] is a syntactic class of programs that enjoy the *sch-property*, for any extensional database [10]. Its programs are characterized through a body variable *marking procedure* whose input is the set \mathcal{P}^r of program rules (the data do not participate).

The procedure has two steps: (a) *Preliminary step*, for each $\sigma \in \mathcal{P}^r$ and variable $X \in \text{body}(\sigma)$, if there is an atom $A \in \text{head}(\sigma)$ where X does not appear, mark each occurrence of X in $\text{body}(\sigma)$, and (b) *Propagation step*, for each $\sigma \in \mathcal{P}^r$, if a marked variable in $\text{body}(\sigma)$ appears at position π , then for every $\sigma' \in \mathcal{P}^r$ (including σ), mark each occurrence of the variables in $\text{body}(\sigma')$ that appear in $\text{head}(\sigma')$ in the same position π .

\mathcal{P}^r is *sticky* when, after applying the marking procedure, there is no rule with a marked variable appearing more than once in its body (notice that a variable never appears both marked and unmarked in a same body).

Example 5. The initial set of three rules, \mathcal{P}^r , is shown on the left-hand side below. The second rule already shows marked variables (with hat) after the preliminary step. The set of rules on the right-hand side are the result of whole marking procedure.

$$\begin{array}{ll} r(X, Y), p(X, Z) \rightarrow s(X, Y, Z). & r(\hat{X}, Y), p(\hat{X}, \hat{Z}) \rightarrow s(X, Y, Z). \\ s(\hat{X}, Y, \hat{Z}) \rightarrow u(Y). & s(\hat{X}, Y, \hat{Z}) \rightarrow u(Y). \\ u(X) \rightarrow \exists Y r(Y, X). & u(X) \rightarrow \exists Y r(Y, X). \end{array}$$

Variables X and Z in the first rule-body end up marked after the propagation step: they appear in the same rule's head, in marked positions ($s[1]$ and $s[3]$ in the body of the second rule). Accordingly, the set of rules is *not* sticky: X in the first rule's body is marked and occurs twice (in $r[1]$ and $p[1]$). ■

With sticky programs, QA can be done in polynomial-time in data complexity [10]. A program with the *sch-property* may not be syntactically sticky. Actually, the *SCh* class can be extended to several larger, semantic, classes of programs that enjoy a form of the *sch-property* with the propagation condition during the chase only on values in certain forms of “infinite” positions. (We propose a new, syntactic class along these lines in Section 4). Something similar can be done with the class of sticky programs.

Weakly-Sticky (WS) Programs. This is a syntactic class that extends those of *WA* and sticky programs. Its characterization uses the above notions of finite-rank and marked variable: A set of rules \mathcal{P}^r is *WS* if, for every rule in it and every repeated variable in its body, the variable is either non-marked or appears in some position in $\Pi_F(\mathcal{P}^r)$.

Example 6. (example 5 cont.) \mathcal{P}^r is *WS*, because $p[1] \in \Pi_F(\mathcal{P}^r)$; and X , the only repeated variable in a body (of the first rule), is marked, but in $p[1]$. ■

The *WS* condition guarantees tractability of QA, because CQs can be answered on an initial fragment of the chase whose size is polynomial in that of the extensional database. This relies on these facts: (a) Finite-rank positions can be saturated by polynomially many values in the size of the extensional database. (b) Stickiness for infinite-rank positions ensures that polynomially many values

are required in them for answering a query at hand. In fact, stickiness for infinite positions makes the number of values required in them for QA polynomially depend on the number of values in finite-rank positions. So, both in finite and infinite-rank positions, polynomially many values are needed.

The above argument about QA is more general than as applied to *WS* programs. It can be applied with more general, syntactic and semantic, classes of programs that are characterized through the use of the stickiness condition on positions where infinitely many values may appear during the chase. *WS* programs are a special case, where those positions are with infinite-rank; and the stickiness is enforced by the syntactic variable-marking mechanism. Actually, we can make the general claim that the combination of finitely many values in finite positions plus chase-stickiness on infinite positions makes QA decidable.

Generalized Stickiness. The *generalized-stickiness of the chase* (*gsch-property*) is defined by relaxing the condition in the *sch-property*: the condition applies to values for the repeated body variables that do not appear in *finite positions*. *GSCh* is the semantic class of programs with the *gsch-property* (cf. Figure 1).

Example 7. (ex. 4 cont.) \mathcal{P}_1 and \mathcal{P}_2 have no infinite positions because for both programs the chase terminates. Consequently, they are *GSCh*. Consider a program \mathcal{P}_3 with the same database $D_3 = D_1$ and a set \mathcal{P}_3^r of rules which is $\mathcal{P}_2^r \cup \{\sigma\}$ such that, $\sigma: r(X, Y) \rightarrow \exists Z r(Z, X)$. $r[1]$ and $r[2]$ are infinite positions because, during the chase of \mathcal{P}_3 , σ cyclically generates infinite null values in $r[2]$ that also propagate to $r[1]$. The chase of \mathcal{P}_3 does not have the *gsch-property* and it is not *GSCh* since the value b replaces the repeated body variable Y that only appears in infinite positions ($r[1]$ and $r[2]$) and b does not propagate all the way down during the chase procedure. ■

4 Selection Functions and Program Classes

The finite positions in the definition of the *gsch-property* are not computable for a given program which makes it impossible to decide if the program has the property. Here, we define selection functions that determine subsets of the finite positions of a program. We replace finite positions in the definition of the *gsch-property* with the results from selection functions in order to define new stickiness properties and program classes.

A *selection function* \mathcal{S} (over a schema \mathcal{R}) is a function that takes a program \mathcal{P} and returns a subset of $\text{FinPoss}(\mathcal{P})$. Particular functions are \mathcal{S}^\perp and \mathcal{S}^\top , that given a program \mathcal{P} , return the empty set and $\text{FinPoss}(\mathcal{P})$, respectively. The latter may not be computable, and depends on the program's data, which is not the case for the former. Π_F also defines a data-independent selection function, $\mathcal{S}^{\text{rank}}$, that returns the finite-rank positions (there are finitely many values in them in the chase of \mathcal{P} , for any data set [10, Lemma 5.1]). A selection function is “syntactically computable” if it only depends on the rules \mathcal{P}^r of a program \mathcal{P} , and we use the notation $\mathcal{S}(\mathcal{P}^r)$.

The *\mathcal{S} -stickiness* is defined by replacing the finite positions in the definition of the *gsch-property* with a selection function \mathcal{S} : The chase of a program \mathcal{P} has the *\mathcal{S} -stickiness property* if the stickiness condition applies only to values replacing

the repeated body variables that do not appear in a position of $\mathcal{S}(\mathcal{P})$. $SCh(\mathcal{S})$ is the semantic class of programs with the \mathcal{S} -stickiness. In particular, $SCh = SCh(\mathcal{S}^\perp)$, $GSCh = SCh(\mathcal{S}^\top)$. Also, $WSCh = SCh(\mathcal{S}^{rank})$ is the class of programs with *weak-stickiness of the chase*. $SCh(\mathcal{S})$ specifies a range of semantic classes of programs starting with SCh , ending with $GSCh$, and with $WSCh$ in between.

$SCh(\mathcal{S})$ grows monotonically with \mathcal{S} : For selection functions \mathcal{S}_1 and \mathcal{S}_2 over schema \mathcal{R} , if $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $SCh(\mathcal{S}_1) \subseteq SCh(\mathcal{S}_2)$. Here, $\mathcal{S}_1 \subseteq \mathcal{S}_2$ if and only if for every program \mathcal{P} , $\mathcal{S}_1(\mathcal{P}) \subseteq \mathcal{S}_2(\mathcal{P})$. In general, the more finite positions are (correctly) identified (and the consequently, the less finite positions are treated as infinite), the more general subclass of $GSCh$ that is identified or characterized.

Sticky Datalog $^\pm$ uses the marking procedure to restrict the repeated body variables and impose the *sch-property*. Applying this syntactic restriction only on body variables specified by syntactic selection functions results in syntactic classes that extend sticky Datalog $^\pm$. These syntactic classes are subsumed by the semantic classes defined by the same selection functions; each of these syntactic classes only partially represents its corresponding semantic class. Particularly, SCh subsumes sticky Datalog $^\pm$ [10]; and WS is a syntactic subclass of $WSCh$ (cf. (g) and (h) in Figure 1).

5 Joint-Weakly-Sticky Programs

The definition of the class of *JWS* programs uses the syntactic selection function \mathcal{S}^\exists , which appeals to the *existential dependency graph* of a program [15] (to define *joint-acyclic* programs). We briefly review it here.

Let \mathcal{P}^r be a set of rules that is standardized apart, i.e. no variable appears in more than one rule. For a variable X , let $B(X)$ ($H(X)$) be the set of all positions where X occurs in the body (head) of its rule σ . For a \exists -variable Z , the set of target positions of Z , denoted by $T(Z)$, is the smallest set of positions such that (a) $H(Z) \subseteq T(Z)$, and (b) $H(X) \subseteq T(Z)$ for every \forall -variable X with $B(X) \subseteq T(Z)$. Roughly speaking, $T(Z)$ is the set of positions where the null values invented by Z may appear in during the chase.

An *existential dependency graph* (*EDG*) of \mathcal{P}^r is a directed graph with the \exists -variables of \mathcal{P}^r as its nodes. There is an edge from Z to Z' if there exists a body variable X in the rule containing Z' such that $B(X) \subseteq T(Z)$. Intuitively, the edge shows that the values invented by Z might appear in the body of the rule of Z' and cause invention of values by Z' . Therefore, a cycle represents the possibility of infinite null values invention by the \exists -variables in the cycle.

Example 8. Let \mathcal{P}^r contain the following rules: $u(Y), r(X, Y) \rightarrow \exists Z \ r(Y, Z)$ and $r(X', Y'), r(Y', Z') \rightarrow p(X', Z')$. For the variable Y , $B(Y) = \{u[1], r[2]\}$, $H(Y) = \{r[1]\}$. Moreover, $T(Z) = \{r[2], p[2]\}$. The *EDG* of \mathcal{P}^r has Z as its node without any edge since $B(X)$ and $B(Y)$ are not subsets of $T(Z)$. \mathcal{P}^r is not *WA*, because $r[1]$ and $r[2]$ have infinite rank. ■

For a set of rules \mathcal{P}^r , we define the set of *finite-existential positions* of \mathcal{P}^r denoted by $\Pi_F^\exists(\mathcal{P}^r)$ as follows: It is the set of positions that are not in the target set of any \exists -variable in a cycle in *EDG*(\mathcal{P}^r). Intuitively, a position in $\Pi_F^\exists(\mathcal{P}^r)$ is not in the target of any \exists -variable that may invent infinite null values.

Proposition 1. For every set of rules \mathcal{P}^r , $\Pi_F(\mathcal{P}^r) \subseteq \Pi_F^\exists(\mathcal{P}^r)$. \blacksquare

Π_F^\exists defines a computable selection function \mathcal{S}^\exists that returns finite-existential positions of a program (cf. (c) in Figure 1). $SCh(\mathcal{S}^\exists)$ is a new semantic subclass of $GSCh$ that generalizes $SCh(\mathcal{S}^{rank})$ since \mathcal{S}^\exists provides a finer mechanism for capturing finite positions in comparison with \mathcal{S}^{rank} (cf. (e) and (f) in Figure 1).

A program \mathcal{P} is *joint-weakly-sticky* (*JWS*) if for every rule in \mathcal{P}^r and every variable in its body that occurs more than once, the variable is either non-marked or appears in some positions in $\Pi_F^\exists(\mathcal{P}^r)$. The class of *JWS* programs is a proper subset of $SCh(\mathcal{S}^\exists)$ and extends *WS* (cf. (i) and (k) in Figure 1). Specifically, the program in Example 8 is *JWS*, because every position is finite-existential, but not *WS*, because Y' is marked and appears in $r[1]$ and $r[2]$ with infinite rank.

6 A Chase-Based Query Answering Algorithm

SChQA is a QA algorithm for programs in the semantic class of $SCh(\mathcal{S})$. It is based on a bottom-up data generation approach and applies a query-driven chase. The algorithm takes as input a computable selection function \mathcal{S} , a program $\mathcal{P} \in SCh(\mathcal{S})$, and a CQ \mathcal{Q} over schema \mathcal{R} and returns $ans(\mathcal{Q}, \mathcal{P})$.

Before describing SChQA, we introduce some notations. A *homomorphism* is a structure-preserving mapping, $h: \mathcal{C} \cup \mathcal{N} \rightarrow \mathcal{C} \cup \mathcal{N}$, between two instances over schema \mathcal{R} that is the identity on constants. An *isomorphism* is a bijective homomorphism.

Definition 1. A rule $\sigma \in \mathcal{P}^r$ and an assignment θ are *applicable* over an instance I of \mathcal{R} if: (a) $I \models (body(\sigma))[\theta]$; and (b) there is an assignment θ' that extends θ , maps the \exists -variables of σ into fresh nulls, and $\theta'(head(\sigma))$ is *not isomorphic* to any atom in I . \blacksquare

Note that for an instance I and a set of rules \mathcal{P}^r , we can systematically compute the applicable pairs of rule-assignment by first finding $\sigma \in \mathcal{P}^r$ for which $body(\sigma)$ is satisfied by I . That gives an assignment θ for which $(body(\sigma))[\theta] \in I$. Then, we construct θ' as specified in Definition 1 and we iterate over atoms in I and we check if they are isomorphic to $\theta'(head(\sigma))$.

In SChQA, we use the notion of *freezing a null value* that is moving it from \mathcal{N} into \mathcal{C} . It may cause new applicable rule-assignment because it changes isomorphic atoms. Considering an instance I , the *resumption* of a step of SChQA is freezing every null in I and continuing the step. Notice that a pair of rule-assignment is applied only once in Step 2. Moreover, if there are more than one applicable pairs, then SChQA chooses the pair that becomes applicable sooner. SChQA is applicable to any Datalog⁺ program and any selection function, and returns sound answers. However, completeness is guaranteed only when applied to programs in $SCh(\mathcal{S})$ with a computable \mathcal{S} .

Example 9. Consider a program \mathcal{P} with $D = \{s(a, b, c), v(b), u(c)\}$, and a BCQ $\mathcal{Q} : p(c, Y) \rightarrow ans_{\mathcal{Q}}$, and a set of rules \mathcal{P}^r containing (the hat signs show the marked variables):

$$\begin{aligned} \sigma_1 : s(\hat{X}, \hat{Y}, \hat{Z}) \rightarrow \exists W \ s(Y, Z, W). \quad \sigma_2 : u(\hat{X}) \rightarrow \exists Y, Z \ s(X, Y, Z). \\ \sigma_3 : s(\hat{X}, Y, Z), v(\hat{X}), s(Y, Z, \hat{W}) \rightarrow p(Y, Z). \end{aligned}$$

Algorithm 1 The SChQA algorithm

Inputs: A selection function \mathcal{S} , a program $\mathcal{P} \in \text{SCh}(\mathcal{S})$, and a CQ \mathcal{Q} over \mathcal{P} .

Output: $\text{ans}(\mathcal{Q}, \mathcal{P})$.

Step 1: Initialize an instance I with the extensional database D .

Step 2: Choose an applicable rule-assignment σ and θ over I , add $\text{head}(\sigma)[\theta']$ into I in which θ' is an extension of θ with mappings for the \exists -variables in σ to fresh nulls in \mathcal{N} .

Step 3: Freeze the nulls in the new atom in Step 2 that appear in the positions of $\mathcal{S}(\mathcal{P})$.

Step 4: Iteratively apply Steps 2 and 3 until there is no more applicable pair of rule-assignment.

Step 5: Resume Step 2 with I , i.e. freeze nulls in I and continue with Steps 2. Repeat resumption $M_{\mathcal{Q}}$ times where $M_{\mathcal{Q}}$ is the number of variables in \mathcal{Q} .

Step 6: Return the tuples in $\mathcal{Q}(I)$ that do not have null values (including the frozen nulls).

\mathcal{P} is in WS and so $\text{SCh}(\mathcal{S}^{\text{rank}})$. Specifically in σ_3 , X occurs in $v[1]$ which is in $\mathcal{S}^{\text{rank}}(\mathcal{P}^r)$ and Y and Z are not marked. The algorithm starts from $I = D$. At Step 2, σ_1 and $\theta_1 = \{X \rightarrow a, Y \rightarrow b, Z \rightarrow c\}$ are applicable; and SChQA adds $s(b, c, \zeta_1)$ into I . σ_2 and $\theta_2 = \{X \rightarrow c\}$ are also applicable and they add $s(c, \zeta_2, \zeta_3)$ into I . Note that Step 3 does not freeze ζ_1 , ζ_2 , and ζ_3 since they are not in $\mathcal{S}^{\text{rank}}(\mathcal{P}^r)$.

There is not more applicable rule-assignments and we continue with Step 5. Consider that σ_1 and $\theta_3 = \{X \rightarrow b, Y \rightarrow c, Y \rightarrow \zeta_1\}$ are not applicable since any $\theta'_3 = \theta_3 \cup \{W \rightarrow \zeta_4\}$ generates $s(c, \zeta_1, \zeta_4)$ that is isomorphic with $s(c, \zeta_2, \zeta_3)$ already in I . SChQA is resumed once since \mathcal{Q} has one variable. This is done by freezing $\zeta_1, \zeta_2, \zeta_3$ and returning to Step 2. Now, $s(c, \zeta_1, \zeta_4)$ and $s(c, \zeta_2, \zeta_3)$ are not isomorphic anymore and σ_1 and θ_3 are applied which results in $s(c, \zeta_1, \zeta_4)$. As a consequence, σ_3 and $\theta_4 = \{X \rightarrow b, Y \rightarrow c, Z \rightarrow \zeta_1, W \rightarrow \zeta_4\}$ are applicable, which generate $p(c, \zeta_1)$. The instance I in Step 6 is $I = D \cup \{s(b, c, \zeta_1), s(c, \zeta_2, \zeta_3), s(c, \zeta_1, \zeta_4), p(c, \zeta_1), s(\zeta_2, \zeta_3, \zeta_5), s(\zeta_1, \zeta_4, \zeta_6)\}$, and $I \models \mathcal{Q}$. ■

The number of resumptions with SChQA depends on the query. However, for practical purposes, we could run SChQA with N resumptions, to be able to answer queries with up to N variables. If a query has more than N variables, we can incrementally retake the already-computed instance I , adding the required number of resumptions.

Theorem 1. Consider a computable selection function \mathcal{S} , a program $\mathcal{P} \in \text{SCh}(\mathcal{S})$, and a CQ \mathcal{Q} over schema \mathcal{R} . Algorithm SChQA taking \mathcal{S} , \mathcal{P} , and \mathcal{Q} as inputs, terminates returning $\text{ans}(\mathcal{Q}, \mathcal{P})$. ■

Termination is due to condition (b) in Definition 1, which prevents isomorphic atoms in I . Note that because of Step 3 the null values that appear in the positions of $\mathcal{S}(\mathcal{P})$ are treated as constants while deciding isomorphic atoms.

However, condition (b) in Definition 1 prevents some atoms from I that are necessary for answering \mathcal{Q} . Adding these atoms depends on the applicability of certain pairs of rule-assignment in which the assignment replaces some repeated variables in the body of the rule with null values. Each resumption makes some of these pairs applicable by freezing nulls. Since \mathcal{P} is $SCh(\mathcal{S})$, there are at most $M_{\mathcal{Q}}$ such rules and so $M_{\mathcal{Q}}$ resumptions are sufficient for answering \mathcal{Q} . The running time of SChQA depends on the number of finite values that may appear in the positions of $\mathcal{S}(\mathcal{P})$.

Proposition 2. Algorithm SChQA runs in PTIME in data if the following holds for \mathcal{S} : for any program \mathcal{P}' , the number of values appearing in $\mathcal{S}(\mathcal{P}')$ -positions during the chase is polynomial in the size of the extensional data. ■

Lemma 1. During the chase of a Datalog⁺ program \mathcal{P} , the number of distinct values in $\mathcal{S}^{\exists}(\mathcal{P}^r)$ -positions is polynomial in the size of the extensional data. ■

Corollary 1. SChQA runs in PTIME in data with programs in $SCh(\mathcal{S}^{\exists})$, in particular for the programs in the *JWS* and *WS* syntactic classes. ■

7 Magic-Sets and *JWS* Datalog[±]

Magic-sets is a general technique for rewriting logical rules so that they may be implemented bottom-up in a way that avoids the generation of irrelevant facts [7, 11]. The advantage of such a rewriting technique is that, by working bottom-up, we can take advantage of the structure of the query and the data values in it, optimizing the data generation process.

In this section, we present a magic-sets rewriting for Datalog⁺ programs, denoted by MagicD⁺. It has two changes regarding the technique in [11] in order to: (a) work with \exists -variables in the existential rules, and (b) consider the extensional data of the predicates that also have intensional data defined by the rules. For (a), we apply the solution proposed in [1]. However (b) is specifically relevant for Datalog⁺ programs that allow predicates with both extensional and intentional data, and we address it in MagicD⁺. MagicD⁺ is described in detail in Appendix D in [23].

Example 10. (ex. 8 cont.) Consider a BCQ $\mathcal{Q} : p(a, Y) \rightarrow ans_{\mathcal{Q}}$ over a program \mathcal{P} with $D = \{u(a), r(a, b)\}$ and the rules in \mathcal{P}^r . MagicD⁺ has the following steps:

1. Generate the adorned version of the query by annotating its body predicates with strings of *bs* and *fs* that correspond to the positions with constants or variables respectively. Then, propagate the adorned predicates to the other program rules. Here, $p^{bf}(a, Y) \rightarrow ans_{\mathcal{Q}}$ is the adorned query; $r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z)$ and $u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z)$ are the adorned rules. Note that the first rule in \mathcal{P}^r is not adorned by bounding Z in the head (e.g. $r^{fb}(Y, Z)$) since the \exists -variables can not be bounded.
2. Add magic predicates to the body of the adorned rule. The magic predicates specify the values for the bounded variables: $mg_p^{bf}(X), r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z)$ and $mg_r^{bf}(Y), u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z)$.
3. Generate magic rules that define the magic predicates: $mg_p^{bf}(X) \rightarrow mg_r^{bf}(X)$ and $mg_r^{bf}(X), r^{bf}(X, Y) \rightarrow mg_p^{bf}(a)$, and a fact $mg_p^{bf}(a)$.

4. For the adorned predicates with extensional data (e.g. r), generate new rules to load their extensional data: $mg_{\mathcal{L}^{bf}}(X), r(X, Y) \rightarrow r^{bf}(X, Y)$ and $mg_{\mathcal{L}^{fb}}(Y), r(X, Y) \rightarrow r^{fb}(X, Y)$.

The result is a program \mathcal{P}_m with schema \mathcal{R}_m , $D_m = D$, the set of rules \mathcal{P}_m^r specified in Steps 2-5, and \mathcal{Q}_m which is the adorned query from Step 1. ■

MagicD⁺ differs from the rewriting algorithm of [1] in Step 4. Particularly, in the latter Step 4 is not needed since, unlike the former, it assumes the intentional predicates in \mathcal{P} and the adorned predicates in \mathcal{P}_m do not have extensional data. Therefore, the correctness of **MagicD⁺**, i.e. $ans(\mathcal{Q}, \mathcal{P}) = ans(\mathcal{Q}_m, \mathcal{P}_m)$, follows from both the correctness of the rewriting algorithm in [1] and Step 4.

\mathcal{P}_m^r has certain syntactic properties. First, the magic rules do not have \exists -variables. Also as mentioned in Step 1, the positions of \exists -variables in the head of a rule never become bounded. Additionally we assume that the full information about bounded variables is propagated from the head of an atom to its body. That is when a variable is in a bounded position in the head it appears in the body only in bounded positions.

Applying **MagicD⁺** over a *WS* program \mathcal{P} , \mathcal{P}_m is not necessarily *WS* or in $SCh(\mathcal{S}^{rank})$ (cf. Example 14 in Appendix E in [23]), which means $SCh(\mathcal{S}^{rank})$ and *WS* are not closed under **MagicD⁺**. This is because **MagicD⁺** introduces new join variables between the magic predicates and the adorned predicates, and these variables might be marked and appear only in the infinite rank positions. That means the joins may break the \mathcal{S}^{rank} -stickiness as it happens in Example 14 in Appendix E [23]. Specifically it turned out to be because \mathcal{S}^{rank} decides some finite positions of \mathcal{P}_m^r as infinite rank positions. In fact, the positions of the new join variables are always bounded and are finite. Therefore, **MagicD⁺** does not break \mathcal{S} -stickiness if we consider a finer selection function \mathcal{S} that decides the bounded positions as finite. We show in Theorem 2 that the class of $SCh(\mathcal{S}^\exists)$ and its subclass of *JWS* are closed under **MagicD⁺** since they apply \mathcal{S}^\exists that better specifies finite positions compared to \mathcal{S}^{rank} .

Theorem 2. Let \mathcal{P} and \mathcal{P}_m be the input and the result programs of **MagicD⁺** respectively. If \mathcal{P} is *JWS*, then \mathcal{P}_m is *JWS*. ■

As a result of Theorem 2, we are able to apply **MagicD⁺** in order to optimize *SChQA* for the class of *JWS* and its subclasses sticky and *WS*.

8 Conclusion and Future Research

We introduced semantic and syntactic extensions of sticky and *WS* Datalog[±] and we proposed a practical bottom-up QA algorithm for these programs. We applied a magic-set rewriting technique, **MagicD⁺**, to optimize the QA algorithm. As the future work, we intend to study the applications of the magic-set rewriting for Datalog[±] ontologies and in the presence of program constraints, i.e. negative constraints and equality generating dependencies and specifically for the purpose of managing inconsistency for these ontologies. We believe that *SChQA* and **MagicD⁺** are applicable on real-world scenarios and we plan to implement them and run experiments on real-world data with large data sets.

References

- [1] Alviano, M., Leone, N., Manna, M., Terracina, G. and Veltri, P. Magic-Sets for Datalog with Existential Quantifiers. *Proc. Datalog'12*, 2012, 12:701-718.
- [2] Arenas, M., Gottlob, G. and Pieris, A. Expressive Languages for Querying the Semantic Web. *Proc. PODS*, 2014, pp. 14-26.
- [3] Artale, A., Calvanese, D., Kontchakov, R. and Zakharyashev, M. The DL-Lite Family and Relations. *Journal of Artificial Intelligence*, 36, 2009, pp. 1-69.
- [4] Baget, J. F., Leclère, M., Mugnier, M. L. and Salvat, E. Extending Decidable Cases for Rules with Existential Variables. *Proc. IJCAI*, 2009, 677-682.
- [5] Beeri, C. and Vardi, M. Y. The Implication Problem for Data Dependencies. *Proc. ICALP*, 1981, 1981:73-85.
- [6] Beeri, C. and Vardi, M. Y. A Proof Procedure for Data Dependencies. *Journal of ACM*, 1984, 31(4):718-741.
- [7] Beeri, C. and Ramakrishnan, R. On the Power of Magic. *Proc. PODS*, 1987, 269-284.
- [8] Calì, A., Gottlob, G. and Lukasiewicz, T. A General Datalog-Based Framework for Tractable Query Answering over Ontologies. *Journal of Web Semantics*, 2012, 14:57-83.
- [9] Calì, A., Gottlob, G. and Pieris, A. Ontological Query Answering under Expressive Entity-Relationship Schemata. *Information Systems*, 2012, 37(4):320-335.
- [10] Calì, A., Gottlob, G. and Pieris, A. Towards More Expressive Ontology Languages: The Query Answering Problem. *Artificial Intelligence*, 2012, 193:87-128.
- [11] Ceri, S., Gottlob, G. and Tanca, L. *Logic Programming and Databases*. Springer, 1990.
- [12] Deutsch, A., Nash, A. and Rummel, J. The Chase Revisited. *Proc. PODS*, 2008, pp. 149-158.
- [13] Fagin, R., Kolaitis, P. G., Miller, R. J. and Popa, L. Data Exchange: Semantics and Query Answering. *TCS*, 2005, 336:89-124.
- [14] Johnson, D. S. and Klug, A. Testing Containment of Conjunctive Queries under Functional and Inclusion Dependencies. *Proc. PODS*, 1984, pp. 164-169.
- [15] Krötzsch, M. and Rudolph, S. Extending Decidable Existential Rules by Joining Acyclicity and Guardedness. *Proc. IJCAI*, 2011, pp. 963-968.
- [16] Leone, N., Manna, M., Terracina, G. and Veltri, P. Efficiently Computable Datalog[±] Programs. *Proc. KR*, 2012, pp. 13-23.
- [17] Maier, D., Mendelzon, A. and Sagiv, Y. Testing Implications of Data Dependencies. *Proc. TODS*, 1979, pp. 152-152.
- [18] Marnette, B. Generalized Schema-Mappings: from Termination to Tractability. *Proc. PODS*, 2009, pp. 13-22.
- [19] Meier, M., Schmidt, M. and Lausen, G. Efficiently Computable Datalog[±] Programs. *Proc. VLDB Endowment*, 2009, pp. 970-981.
- [20] Milani, M. and Bertossi, L. Ontology-Based Multidimensional Contexts with Applications to Quality Data Specification and Extraction. *Proc. RuleML*. 2015, pp. 277-293.
- [21] Milani, M., Calì, A. and Bertossi, L. Query Answering on Expressive Datalog[±] Ontologies. *To appear in AMW*. 2016.
- [22] Milani, M., Calì, A. and Bertossi, L. A Hybrid Approach to Query Answering under Expressive Datalog[±]. *Conference submission*. 2016.
- [23] Milani, M. and Bertossi, L. Extending Weakly-Sticky Datalog[±]: Query-Answering Tractability and Optimizations. Extended version of this paper. <https://goo.gl/bJ8MGA>

- [24] Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M. and Rosati, R. Linking Data to Ontologies. *Journal on Data Semantics*. 2008, pp. 133-173.

A Proofs

Proof of Proposition 1: We use proof by contradiction. Assume there is a position π such that: $\pi \in \Pi_F(\mathcal{P}^r)$ and $\pi \notin \Pi_F^\exists(\mathcal{P}^r)$. The latter means there is a cycle in $EDG(\mathcal{P}^r)$ that includes an \exists -variable Z in a rule σ such that $\pi \in T(Z)$. The definition of EDG implies that, there is \forall -variable X in the body of σ for which $B(X) \subseteq T(Z)$. Let π_Z and π_X be the two positions where Z and X appear in σ resp. Then, there is a path from π_Z to π_X and there is also a special edge from π_X to π_Z in $DG(\mathcal{P}^r)$ making a cycle including π_Z with a special edge. Therefore, $\pi_Z \notin \Pi_F(\mathcal{P}^r)$. Since $\pi \in T(Z)$, we can conclude that $\pi \notin \Pi_F(\mathcal{P}^r)$ which contradicts the assumption and completes the proof. ■

Proof of Theorem 1: Let I_i be the instance I after the i -th resumption in SChQA. To prove the termination, we first show that for a finite i there are finitely many terms and so finitely many atoms in I_i . Since the algorithm only adds atoms this suffices to prove the algorithm always stops by reaching a fixed point.

Now, let $f^{\mathcal{P}}$ be the number of terms (constants and nulls) that appear in the positions of $\mathcal{S}(\mathcal{P})$ in I during SChQA, and let $r^{\mathcal{P}}$ and $w^{\mathcal{P}}$ be the number of distinct predicate names and the maximum arity of predicates in \mathcal{P} respectively. Starting from I_0 , since there are no isomorphic atoms in I_0 , there are at most $r^{\mathcal{P}} \times w^{\mathcal{P}}$ nulls (not frozen) and $r^{\mathcal{P}} \times w^{\mathcal{P}} + f^{\mathcal{P}}$ possible terms in I_0 . Considering $r^{\mathcal{P}}, w^{\mathcal{P}}, f^{\mathcal{P}}$ are finite, I_0 is finite. After the first resumption, the $r^{\mathcal{P}} \times w^{\mathcal{P}}$ nulls are frozen; and at most $r^{\mathcal{P}} \times w^{\mathcal{P}}$ new nulls are invented. Now in I_1 , there are at most $2 \times r^{\mathcal{P}} \times w^{\mathcal{P}} + f^{\mathcal{P}}$ terms which means I_1 is also finite. With the same line of reasoning, we can prove that I_i with a finite i has finite terms, $i \times r^{\mathcal{P}} \times w^{\mathcal{P}} + f^{\mathcal{P}}$, and it is finite. Since there are M_Q resumptions and M_Q is finite, SChQA terminates.

SChQA is sound because Step 2 is sound and it only adds atoms into I that are entailed by the rules in \mathcal{P}^r .

For the proof of completeness, we assume Q is a BCQ. Note that for free CQs we can make a BCQ for every tuple in the answers set and apply the same proof for the obtained BCQs. To prove the completeness of SChQA, i.e. $\mathcal{P} \models Q \Rightarrow I_{M_Q} \models Q$, it is enough to show $I_\infty \models Q \Rightarrow I_{M_Q} \models Q$. That is because I_∞ (the instance after infinitely many resumptions) gives the same answers that are obtained from the chase of \mathcal{P} , since every null value in I_∞ is eventually frozen and condition (b) in Definition 1 is always satisfied.

Let $\mathcal{P} \models Q$ then, as it is proved in [10], there is a *proof-schema* T for Q w.r.t \mathcal{P} . A proof-schema (called *accepting resolution proof-schema* in [10]) is a tree with its nodes and edges labeled with atoms of the schema $\mathcal{R} \cup \{ans_Q\}$ and the rules in $\mathcal{P}^r \cup \{Q\}$ resp. The terms in the atoms are either constants in D or variables. In T , the root node is labeled with ans_Q and there is an assignment θ of the variables in the labels of the nodes in T into the constants in D and nulls that maps the labels of the nodes (other than the root node) into the atoms in $chase(\mathcal{P})$. For every leaf node, h maps its label into an atom in D . The label of

the incoming edges into a node are a rule that shows how the atom of the node is obtained from the atoms in its child nodes. A proof-schema has other syntactic properties that are described in [10, Definition 3.5]. Without loss of generality, we assume that (a) T has minimum height, and (b) θ maps T into the atoms of $\text{chase}(\mathcal{P})$ that are obtained sooner during the chase procedure. In the rest of the proof, whenever we refer to a node as an atom we mean the atom in the label of the node.

Since $I_\infty \models \mathcal{Q}$, there is also an assignment θ' that maps nodes of T into the atoms in I_∞ . The rest of the proof is devoted to show that θ' maps the nodes of T into I_{M_Q} which proves $I_{M_Q} \models \mathcal{Q}$. We do that by showing every variable in T that appears in more than one branch is mapped by θ' into a term that is either a constant or a frozen null in I_{M_Q} .

Let X_1, \dots, X_n be the variables that appear in more than one branch of T and do not occur in any position of $\mathcal{S}(\mathcal{P})$, ordered by the depth they first occur in T (X_1 is the deepest). The \mathcal{S} -stickiness implies that $n \leq M_Q$. That is because these variables represent joins between values that do not appear in the $\mathcal{S}(\mathcal{P})$ positions and so they propagate all the way to the query. Therefore the number of these values (and so the variables) is restricted by the number of variables in the query. Also, let A_1, \dots, A_n be the nodes (atoms) in T where X_1, \dots, X_n first appear. We claim that $\theta'(A_i)$ is in I_{i-1} for each $i \leq n$.

Consider T_1 , the subtree of A_1 . Its leaf nodes are mapped by θ' into $D \subseteq I_0$ according to the definition of T . In the internal nodes, if a variable appears in more than one branch of T_1 , it occurs at least once in a position of $\mathcal{S}(\mathcal{P})$ in each branch. Now consider the variable Y as the first such variable and A_Y as the node where it first appears in T_1 and A'_Y as the node where the branches meet. A_Y is in I_0 because of the assumptions (a) and (b). Additionally, if the term $t = h'(Y)$ is a null value, it is frozen in I_0 because it appears in some positions of $\mathcal{S}(\mathcal{P})$. Note that even if Y occurs in A_Y in a non- $\mathcal{S}(\mathcal{P})$ position which means t is not frozen immediately in Step 3 of SChQA, t will eventually become frozen in I_0 before reaching A'_Y . That is because any other isomorphic atom B with the term t' that prevented A_Y from I_0 (according to condition (b) in Definition 1) will eventually propagate to the same position π and becomes frozen and will not be isomorphic to A_Y anymore. Note that we assumed there is no other join, so if t was going to propagate to A'_Y , t' will also propagate to A'_Y and its \mathcal{S} -finite position. Similarly, we can prove that every variable that appears in more than one branch of T_1 is mapped by θ' into a term that is either constant or is frozen in I_0 . Therefore every term in the atoms of the subtree T_1 are frozen in I_0 and so the nodes in T_1 are mapped by θ' into I_0 .

Now since $\theta'(A_1)$ is in I_0 , the term $\theta'(A_1)$ is frozen in I_1 . Similarly, we can prove that A_2 is in I_1 considering that $\theta'(A_1)$ is frozen in I_1 and continuing with this line of reasoning we can prove that $\theta'(A_i)$ is in I_{i-1} . That means every join variable in T is mapped by θ' into either a constant or a null that is frozen in I_n with $n \leq M_Q$. Therefore T is mapped by θ' into I_{M_Q} which completes our proof of the completeness of SChQA. ■

Proof of Proposition 2: The condition implies that $f^{\mathcal{P}}$ (cf. the proof of Theorem 1) is polynomial w.r.t the extensional data of \mathcal{P} . As a result, the number of terms in I_i (the instance in SChQA after i -th resumption) $i \times r^{\mathcal{P}} \times w^{\mathcal{P}} + f^{\mathcal{P}}$ and also the size of I_i are polynomial in the size of the extensional data. Since the algorithm only adds atoms to the current instance I (never removes atoms from I), that means SChQA stops in PTIME in the size of extensional data. ■

Proof of Lemma 1: The proof is similar to the proof of [13, Theorem 3.9]. The theorem shows the chase of a WA program has polynomial length in the size of the extensional data of the program.

We define \exists -rank of a position π in a predicate in \mathcal{P}^r as the maximum length of a path in $EDG(\mathcal{P})$ ending with Z such that $\pi \in T(Z)$. A finite-existential position has a finite \exists -rank, since it is not in the target of any \exists -variable that is in a cycle in $EDG(\mathcal{P})$.

We prove by induction that: For every finite $i > 0$, there is a polynomial function f_i such that the number of values that appear in the positions with \exists -rank i is at most $f_i(d)$ with $d = \text{size}(D)$.

Base case: The positions with \exists -rank of 0 are not in the target of any \exists -variable. Therefore, these positions can only contain constants from D , and $f_0 = d$.

Inductive step: The values that appear in a position of \exists -rank i are either (a) from the other positions with the same \exists -rank, or (b) from positions with the \exists -rank $j < i$. For (b), they are by inductive hypothesis at most $f_{i-1}(d)$. In case of (a), the values are invented by an \exists -variable Z that is at the end of a path of length i in $EDG(\mathcal{P})$. If there are b_Z variables in the body of the rule of Z , the rule can invent $f_{i-1}(d)^{b_Z}$ new values for the positions with \exists -rank i . There are at most $s_{\mathcal{P}}$ such \exists -variables where $s_{\mathcal{P}}$ is the maximum number of rules in \mathcal{P}^r . Therefore $f_i(d) = s_{\mathcal{P}} \times f_{i-1}(d)^{b_Z} + f_{i-1}(d)$ and since $s_{\mathcal{P}}$ and b_Z are independent of data, f_i is PTIME w.r.t d .

Considering that $i \leq k$ and k (the maximum \exists -rank in \mathcal{P}) is independent of the data of \mathcal{P} , we conclude that $f_k(d)$ is the polynomial maximum number of distinct values in the positions of $\Pi_F^{\exists}(\mathcal{P}^r)$ which proves the proposition. ■

Proof of Theorem 2: To prove \mathcal{P}_m is in $SCh(\mathcal{S}^{\exists})$ we show every repeated variable in \mathcal{P}_m preserves the \mathcal{S}^{\exists} -stickiness property.

First we claim that every bounded position in \mathcal{P}_m is in $\Pi_F^{\exists}(\mathcal{P}_m)$. That is specifically because an \exists -variable never gets bounded during MagicD⁺ and also if a position in the head is bounded the corresponding variable appears in the body only in the bounded positions. As a result, a bounded position can not be in the target of any \exists -variable which proves the claim.

Also note that if a position in \mathcal{P} is finite-existential (the position is in $\Pi_F^{\exists}(\mathcal{P})$), its corresponding position in \mathcal{P}_m is also finite-existential. The prove is by assuming that there is a finite-existential position $\pi \in \mathcal{P}$ and its corresponding position $\pi' \in \mathcal{P}_m$ is not finite-existential which means there is a loop in the EDG of \mathcal{P}_m

including a variable Z' such that $\pi' \in T(Z')$. Then it is easy to show there is also a loop in the *EDG* of \mathcal{P} including a variable Z and $\pi \in T(Z)$ meaning that π is not finite-existential which contradicts the assumption and completes the proof.

Now, we specify four types of joins in \mathcal{P}_m : (a) between the adorned predicates in the adorned rules, (b) between the adorned predicates in the magic rules, (c) between the adorned predicates and the magic predicates in the adorned rules, and (d) between the adorned predicates and the magic predicates in the magic rules.

The joins of Type (a) do not break the \mathcal{S}^\exists -stickiness property since they correspond to join variables in \mathcal{P} . If they were not marked in \mathcal{P} they are still not marked in \mathcal{P}_m and if they were at some finite-existential position the same holds for the variable in \mathcal{P}_m and either way the repeated variable in \mathcal{P}_m preserves the \mathcal{S}^\exists -stickiness property. The joins of Type (b), (c), and (d) also preserve the property since their variables appear in a bounded position and we proved the bounded positions are finite-existential. Therefore every type of joins in \mathcal{P}_m satisfies the \mathcal{S}^\exists -stickiness property and so \mathcal{P}_m is in $SCh(\mathcal{S}^\exists)$.

Note that the same prove holds for *JWS* programs, while it does not apply to *WS* and $SCh(\mathcal{S}^{rank})$. The latter because the two claims at the beginning of the proof does not hold for these programs. ■

B The Chase Procedure

The chase procedure of a program \mathcal{P} with database D and rules \mathcal{P}^r starts from the extensional database D and it iteratively applies the rules in \mathcal{P}^r through some chase steps. In a chase step, the procedure applies a rule $\sigma \in \mathcal{P}^r$ and an assignment θ on the current instance I . σ and θ are applicable if θ maps the body of σ into I . Let θ' be an extension of θ that maps the \exists -variables of σ into fresh nulls in \mathcal{N} . The result of applying σ and θ over I is an instance $I' = I \cup \{\theta'(\text{head}(\sigma))\}$. We denote a chase step by $I \xrightarrow{\sigma, \theta} I'$.

Based on chase steps, the *level* of an atom is defined as follows: For an atom $a \in D$, $\text{level}(a) = 0$. If an atom is the result of a chase step, $I_{i-1} \xrightarrow{\sigma_i, \theta_i} I_i$, let $\text{level}(a) = \max_{b \in \theta_i(\text{body}(\sigma_i))} (\text{level}(b) + 1)$. We refer to the chase with atoms up to level k as $\text{chase}^k(\mathcal{P})$, while $\text{chase}^{[k]}(\mathcal{P})$ is the instance constructed after $k \geq 0$ chase steps.

Note that, the chase steps are applied in a *level saturating* fashion, meaning that if there are more than one applicable rules, the one that has body atoms with smallest maximum level is applied. Also importantly, each pair of applicable rule and homomorphism is only applied once during the chase procedure.

The chase procedure stops if there is no applicable rule and assignment. The chase result, $\text{chase}(\mathcal{P})$ or $\text{chase}(D, \mathcal{P}^r)$ called the chase, is the result of the last chase step. If the chase procedure does not terminate, $\text{chase}(\mathcal{P}) = \bigcup_{i=0}^{\infty} (I_i)$, in which, $I_0 = D$, and, I_i is the result of the i -th chase step for $i > 0$.

C Stickiness Property and its Generalization

In this section, we first formalize the *sch-property* introduced in [10] and we give an extension of it, *generalized stickiness property of the chase* (*gsch-property*). Both the *sch-property* and the *gsch-property* are defined based on the notions of the *chase relation* and the *chase derivation relation* that we explain here.

Definition 2. Let $I_i \xrightarrow{\sigma_i, \theta_i} I_i \cup \{A_i\}$ be the i -th chase step of a program \mathcal{P} that applies the rule σ_i with θ_i as the assignment that makes the body of σ_i true in I_i and generates a new atom A_i . We define $rchase(\mathcal{P}) = \bigcup_{i=1}^M (\sigma_i[\theta_i] \times A_i)$ as the *chase relation* of \mathcal{P} , where M is the minimum number of steps to make the chase stop (but $M = \infty$ if the latter does not stop). The *chase derivation relation* of \mathcal{P} , denoted by $dchase(\mathcal{P})$, is the transitive closure of $rchase(\mathcal{P})$. ■

Intuitively, $dchase(\mathcal{P})$ contains every derivation of atoms in $chase(\mathcal{P})$. In Example 2, $dchase(\mathcal{P})$ includes $(r(a, b), r(b, \zeta_1))$, $(r(a, b), s(a, b, \zeta_1))$ and $(r(a, b), r(\zeta_1, \zeta_2))$.

Definition 3. A program \mathcal{P} has the *stickiness property* of the chase [10], the *sch-property*, if and only if for every chase step $I_i \xrightarrow{\sigma_i, \theta_i} I_i \cup \{A_i\}$, the following holds: If a variable X appears more than once in $body(\sigma_i)$, $\theta_i(X)$ occurs in A_i and every atom B for which, $(A_i, B) \in dchase(\mathcal{P})$. SCh is the class of programs with the *sch-property*. ■

The concept of the *gsch-property* is specified by relaxing the condition for the *sch-property*: it applies only to values for repeated variables in the body of σ_i that do not appear in so-called *finite positions* defined next.

Definition 4. Given a program \mathcal{P} with schema \mathcal{R} , the set of finite positions of \mathcal{P} , referred to as $FinPoss(\mathcal{P})$, is the set of positions where finitely many values appear in $chase(\mathcal{P})$. Every position that is not finite is infinite. ■

Definition 5. A program \mathcal{P} has the *generalized-stickiness property of the chase* (*gsch-property*) if and only if for every chase step, $I_i \xrightarrow{\sigma_i, \theta_i} I_i \cup \{A_i\}$, the following holds: If a variable X appears more than once in $body(\sigma_i)$ and *not* in $FinPoss(\mathcal{P})$, $\theta_i(X)$ occurs in A_i and every atom B for which, $(A_i, B) \in dchase(\mathcal{P})$. GSC is the class of programs with the *gsch-property*. ■

D MagicD⁺

The MagicD⁺ rewriting technique takes a Datalog⁺ program \mathcal{P} and a CQ \mathcal{Q} of schema \mathcal{R} and returns a program \mathcal{P}_m and a CQ \mathcal{Q}_m of schema \mathcal{R}_m such that $ans_{\mathcal{Q}}(\mathcal{Q}, \mathcal{P}) = ans_{\mathcal{Q}_m}(\mathcal{Q}_m, \mathcal{P}_m)$. Here we describe MagicD⁺ in more details using the same program in Example 10.

The rewriting uses the notion of *sideways information passing strategy* (*SIPS*). A *SIPS* of a rule specifies a propagation strategy in a top-down evaluation approach for the rule. Intuitively, a *SIPS* of a rule is a strict partial order over the atoms of the rule which shows how the bindings are originated from the head and propagated through the body.

Definition 6. Let p be a predicate of arity k . An *adornment* for p is a string $\alpha = \alpha_1 \dots \alpha_k$ defined over the alphabet $\{b, f\}$. The i -th argument of p is considered *bound* if $\alpha_i = b$, or *free* if $\alpha_i = f$, ($1 \leq i \leq k$). The predicate p^α is an *adorned predicate* of p . Consider a Datalog⁺ rule σ with a head predicate p and an adornment α of p . Let $atoms(\sigma)$ be the set of atoms in the body and the head of σ . A *SIPS* of σ and α is a pair $\langle <^{\sigma, \alpha}, f^{\sigma, \alpha} \rangle$ in which $<^{\sigma, \alpha}$ is a strict partial order over $atoms(\sigma)$ and $f^{\sigma, \alpha}$ is a function assigning to each atom $A \in atoms(\sigma)$ an adornment such that $<^{\sigma, \alpha}$ and $f^{\sigma, \alpha}$ have the following properties:

1. For every atom $A \in body(\sigma)$, $head(\sigma) <^{\sigma, \alpha} A$.
2. $f^{\sigma, \alpha}(head(\sigma)) = \alpha$.
3. If a variable X in A is bounded according to $f^{\sigma, \alpha}(A)$, X either appears in $head(\sigma)$ and it is bounded according to $f^{\sigma, \alpha}(head(\sigma))$ or it occurs in a body atom $B \in body(\sigma)$ such that $B <^{\sigma, \alpha} A$. Intuitively, this property says if a variable is bounded in an atom it is either bound in the head atom or it is already evaluated in a body atom.

A *SIPS* $\langle <_1^{\sigma, \alpha}, f_1^{\sigma, \alpha} \rangle$ is included in a *SIPS* $\langle <_2^{\sigma, \alpha}, f_2^{\sigma, \alpha} \rangle$ iff for every atom $A \in \sigma$ and variable $X \in A$, if A is bounded according to $f_1^{\sigma, \alpha}(A)$ it is bounded according to $f_2^{\sigma, \alpha}(A)$. A *SIPS* is *partial* if it is included in another *SIPS* and otherwise it is *full*. ■

Intuitively, a *SIPS* is partial if it does not always propagate all available information. In Section 7 and specifically in Theorem 2, we consider full *SIPS*. We discuss about MagicD⁺ with a partial *SIPS* in Section F.

Example 11. (ex. 10 cont.) For the rule $\sigma : r(X, Y), r(Y, Z) \rightarrow p(X, Z)$ and the adornment $\alpha = bf$, a possible *SIPS* is $<^{\sigma, bf} = \{(p(X, Z), r(X, Y)), (p(X, Z), r(Y, Z)), (r(X, Y), r(Y, Z))\}$ and $f^{\sigma, bf} = \{(p(X, Z), bf), (r(X, Y), bf), (r(Y, Z), bf)\}$.

This *SIPS* is complete. A possible partial *SIPS* for σ and α is: $<_{par}^{\sigma, bf} = <_{par}^{\sigma, bf}$ and $f_{par}^{\sigma, bf} = \{(p(X, Z), bf), (r(X, Y), bf), (r(Y, Z), ff)\}$ in which for $f_{par}^{\sigma, bf}(r(Y, Z))$ both positions are free unlike $f^{\sigma, bf}(r(Y, Z))$ with the first position bounded. ■

MagicD⁺ starts from the body atoms of Q and generates their adorned atoms by annotating their predicates with strings of b 's and f 's in the positions that contain constants and variables resp. We make a set of predicates P with two types of adorned predicates: marked and unmarked. We add the new predicates of Q into P as unmarked predicates. Then we iteratively pick an unmarked predicate p^α from P and generate its adorned rules and mark it as processed. For p^α , we find every rule σ with the head predicate p and we generate an adorned

rule σ' as follows. We choose a *SIPS* of σ and α and we replace every body atom in σ with its adorned atom and the head of σ with p^α . The adornment of the body atoms is obtained from the *SIPS* and its function $f^{\sigma, bf}$. If the generated adorned predicates from the body of σ are not in P we add them into P as unmarked predicates. We add the adorned rule σ' into \mathcal{P}^r and after repeating this for every rule σ we mark p .

Example 12. (ex. 10 cont.) For the CQ $p(a, Y) \rightarrow ans_Q$, its adorned rule is $p^{bf}(a, Y) \rightarrow ans_Q$ which adds p^{bf} to P . Adorning $r(X, Y), r(Y, Z) \rightarrow p(X, Z)$ with the head predicate p^{bf} results into an adorned rule $r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z)$ that we add into \mathcal{P}_m^r . We add r^{bf} to P and mark p^{bf} as processed. Next, r^{bf} results into the adorned rule $u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z)$ and adds r^{fb} into P and marks r^{bf} as processed. But, there is no adorned rule for r^{fb} since $u(Y), r(X, Y) \rightarrow \exists Z r(Y, Z)$ can not be bounded in the position of the variable Z . The result set of adorned rule is:

$$r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z). \quad u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z). \quad \blacksquare$$

Now, for every adorned rule σ' in \mathcal{P}^r with the adorned head predicate p^α , we add to the body of σ' a magic atom with predicate $m_{-}p^\alpha$. The arity of $m_{-}p^\alpha$ is the number of occurrences of b in the adornment α , and its variables correspond to the bound variables of head atom of p^α .

The magic predicates are defined by the magic rules constructed as follows. For every occurrence of an adorned predicate p^α in an adorned rule σ' , we construct a magic rule σ'' that defines $mg_{-}p^\alpha$ (a magic predicate might have more than one definition). We assume that the atoms in σ' are ordered according to the partial order in the *SIPS* of σ and α . If the occurrence of p^α is in atom A and there are A_1, \dots, A_n on the left hand side of A in σ' , the body of σ'' contains A_1, \dots, A_n and the magic atom of A in the head. We also create a seed for the magic predicates, in the form of a fact, obtained from the query.

Example 13. (ex. 10 cont.) Adding the magic atom $mg_{-}p^{bf}$ to the adorned rule $r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z)$ we obtain $mg_{-}p^{bf}(X), r^{bf}(X, Y), r^{bf}(Y, Z) \rightarrow p^{bf}(X, Z)$. Similarly the adorned rule $u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z)$ becomes $mg_{-}r^{bf}(Y), u(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z)$. The following are the magic rules that define $mg_{-}p^{bf}$ and $mg_{-}r^{bf}$ (the seed atom for the magic predicates is, $mg_{-}p^{bf}(a)$):

$$mg_{-}p^{bf}(X) \rightarrow mg_{-}r^{bf}(X). \quad mg_{-}r^{bf}(X), r^{bf}(X, Y) \rightarrow mg_{-}r^{bf}(Y). \quad \blacksquare$$

\mathcal{P} is a Datalog⁺ program that might have intentional predicates with extensional data in D . Therefore, we add rules to load the data from D when such a predicate gets adorned. In the Example 10, r is an intentional predicates with the extensional data $r(a, b)$ and so we add the following to load this data into the adorned predicates $mg_{-}r^{bf}$ and $mg_{-}r^{fb}$:

$$mg_{-}r^{bf}(X), r(X) \rightarrow r^{bf}(X). \quad mg_{-}r^{fb}(X), r(X) \rightarrow r^{fb}(X).$$

E Examples

Example 14. Consider a program \mathcal{P} with $D = \{r(a, b)\}$ and the following rules:

$$r(X, Y) \rightarrow \exists Z r(Y, Z). \quad (3)$$

$$c(X), r(X, Y), r(Y, Z) \rightarrow u(X, Z). \quad (4)$$

\mathcal{P} is not *WS* because Y in (4) is marked and does not appear in $\Pi_F(\mathcal{P}^r)$. The program is *WSCh* because (4) is never applied during the chase of \mathcal{P} . ■

Example 15. Consider a program \mathcal{P} with a database $D = \{r(a, b), v(b)\}$, a BCQ $\mathcal{Q} : r(Y, a) \rightarrow \text{ans}_{\mathcal{Q}}$ and the following set of rules \mathcal{P}^r :

$$r(X, Y) \rightarrow \exists Z r(Y, Z). \quad (5)$$

$$r(X, Y) \rightarrow \exists Z r(Z, X). \quad (6)$$

$$r(X, Y), r(Y, Z), v(Y) \rightarrow r(Y, X). \quad (7)$$

The program is *WS* since the only repeated marked variable is Y in (7) and it appears in $v[1] \in \Pi_F(\mathcal{P}^r)$. The marked variables are specified by a hat sign. The result of the magic-sets rewriting \mathcal{P}^m is the following, with the adorned rules:

$$r^{fb}(Y, a) \rightarrow \text{ans}_{\mathcal{Q}}. \quad (8)$$

$$mg_r(Y), r^{fb}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z). \quad (9)$$

$$mg_r(X), r^{bf}(X, Y) \rightarrow \exists Z r^{fb}(Z, X). \quad (10)$$

$$mg_r(X), r^{bf}(X, Y), r^{bf}(Y, Z), v(Y) \rightarrow r^{fb}(Y, X). \quad (11)$$

$$mg_r(Y), r^{fb}(X, Y), r^{bf}(Y, Z), v(Y) \rightarrow r^{bf}(Y, X). \quad (12)$$

and the magic rules:

$$mg_r(a). \quad (13)$$

$$mg_r(X), r^{bf}(X, Y) \rightarrow mg_r(Y). \quad (14)$$

$$mg_r(Y), r^{fb}(X, Y) \rightarrow mg_r(X). \quad (15)$$

Here, every body variable is marked. Note that according to the description of *MagicD⁺* in Appendix 7, the magic predicates mg_r^{fb} and mg_r^{bf} are equivalent and so we replace them with a single predicates, mg_r .

\mathcal{P}_m is *not WS*, since $r^{fb}[1], r^{fb}[2], r^{bf}[1], r^{bf}[2]$, and $mg_r[1]$ are not in $\Pi_F(\mathcal{P}_m^r)$ so; (9), (10), (14) break the syntactic property of *WS*. Following the chase of \mathcal{P}_m , the program is not in *SCh(S^{rank})* either. That is because in (14) a replaces X that appears only in infinite rank positions $mg_r[1]$ and $r^{bf}[1]$.

\mathcal{P}_m is *JWS*. That is because, $r^{fb}[2], r^{bf}[1]$ are in $\Pi_F^{\exists}(\mathcal{P}_m^r)$ and every repeated marked variable appears at least once in one of these two positions which means \mathcal{P}_m is *JWS*. Note that both $r^{fb}[2], r^{bf}[1]$ are bounded positions and are finite-existential which confirms the first claim in the proof of Theorem 2. ■

Example 16. In Example 15, we applied full *SIPS*s, that passe full information about the bounded variables during the evaluation of a rule. Here, we consider partial *SIPS*s that generate the following program:

$$r^{fb}(Y, a) \rightarrow ans_{\mathcal{Q}}. \quad (16)$$

$$mg_{\neg}^{ff}, r^{ff}(X, Y) \rightarrow \exists Z r^{ff}(Y, Z). \quad (17)$$

$$mg_{\neg}r(Y), r^{ff}(X, Y) \rightarrow \exists Z r^{bf}(Y, Z). \quad (18)$$

$$mg_{\neg}r(X), r^{ff}(X, Y) \rightarrow \exists Z r^{fb}(Z, X). \quad (19)$$

$$mg_{\neg}r(X), r^{bf}(X, Y), r^{bf}(Y, Z), v(Y) \rightarrow r^{fb}(Y, X). \quad (20)$$

and the magic rules:

$$mg_{\neg}r(a). \quad (21)$$

$$mg_{\neg}r(X), r^{bf}(X, Y) \rightarrow mg_{\neg}r(Y). \quad (22)$$

$$mg_{\neg}r(Y), r^{fb}(X, Y) \rightarrow mg_{\neg}r(X). \quad (23)$$

Specially, in (17)-(19) the information about the bounded variables from the head atom is not used in the body. In (18) and (19), $mg_{\neg}r[1]$ and r^{ff} are infinite positions and if we follow the chase, there are values that replace the join variables Y and X in these rules and the values do not propagate all the way to the head atoms in the next steps. Therefore, the result program is not *GSch*. This shows that using partial *SIPS*, *GSch* or any of its semantic subclasses of $SCh(\mathcal{S})$ are not closed under $MagicD^+$. ■

F Discussion

F.1 Connection with Partial Grounding Approach

A new hybrid approach for QA over *WS* programs is proposed in [21, 22]. In this approach, a given *WS* program is rewritten by partially grounding some variables and transforming the program into a sticky program w.r.t the extensional data. An input CQ then is combined with the result sticky program to obtain a UCQ to be answered directly on the extensional database.

This hybrid approach that combines bottom-up grounding and backward rewriting is a new promising technique for QA. However, it strongly relies on the syntactic properties of the *WS* and sticky programs. The QA algorithm in this paper applies for a range of semantic programs with certain property of their chase instance rather than specific syntactic properties.

F.2 MagicD⁺ with Partial Sideways Information Passing Strategies

In Section 7 and Theorem 2, we assumed that the $MagicD^+$ always uses full *SIPS* for generating the adorned rules and takes advantage of full information about

the bound and free variables of the head atom and the already evaluated atoms in the body. This is specifically necessary to prove the claim that in the result of MagicD^+ every bounded position is finite-existential and so this is required to prove that the class of JWS program is closed under MagicD^+ .

Example 15 in Appendix E shows a situation when using partial SIPSs in MagicD^+ and rewriting a JWS program, the result is not JWS . This is not a syntactic incident because the result is not even $GSch$. It means there is no semantic or syntactic subclass of $GSch$ that uses a proper selection function and is closed under MagicD^+ .

However, this is not problematic for the integration of SChQA and MagicD^+ with partial SIPSs. That is because the SChQA algorithm is still applicable for the result program with a modification in the applicability condition of in Definition 1. Specifically, the adorned rules without their magic predicates for a set of rules that still preserve the stickiness. Therefore, we can ignore the condition (b) in Definition 1 for the magic rules and obtain a new QA algorithm that freely propagates the data of the magic predicates. With this modification the result algorithm still terminates since the magic predicates do not invent new values and it returns correct answers.

F.3 Further Generalization of the Stickiness Property of the Chase

In the *gsch-property*, we generalized the stickiness by relaxing the condition on the join variables when they appear at least once in a finite position. $GSch$ (the class of program with the *gsch-property*) is an abstract class that can not be syntactically checked but it is an important class as it defines a decidability paradigm for the programs with different form of stickiness of the chase considering the SChQA algorithm that works for any sub class $SCh(\mathcal{S})$ of $GSch$ with computable \mathcal{S} .

Investigating SChQA and the proof of its correctness in Theorem 1, we notice that we could even further relax the stickiness condition and define a more general class compared to $GSch$. That is if a variable is replaced during the chase with a value that traversed a finite position at some point before reaching the current rule, we can relax the condition for the variable. This defines a more general class of programs compared to $GSch$ that still can define decidable semantic classes with computable selection functions.